

# TD5 de Statistique Inférentielle

Modèle linéaire et applications

Magistère d'économiste statisticien

## 1 Cuatro Cuaranta

Soit  $n_1$  et  $n_2$  deux entiers strictement positifs fixés et  $n = 2p > 2 \max(n_1, n_2)$ . On considère pour  $j = 1 \dots n$ , le modèle de régression périodique :

$$Y_j = a_0 + a_1 \cos(2\pi n_1 \frac{j}{n}) + a_2 \cos(2\pi n_2 \frac{j}{n}) + b_1 \sin(2\pi n_1 \frac{j}{n}) + b_2 \sin(2\pi n_2 \frac{j}{n}) + \varepsilon_j.$$

a) Montrer ou admettre pour  $i, i' = 1, 2$ , les relations suivantes :

$$\begin{aligned} \sum_{j=1}^n \cos(2\pi n_i \frac{j}{n}) \cos(2\pi n_{i'} \frac{j}{n}) &= 0 \quad \text{si } i \neq i' \\ &= p \quad \text{si } i = i' \end{aligned}$$

$$\begin{aligned} \sum_{j=1}^n \sin(2\pi n_i \frac{j}{n}) \sin(2\pi n_{i'} \frac{j}{n}) &= 0 \quad \text{si } i \neq i' \\ &= p \quad \text{si } i = i' \end{aligned}$$

$$\sum_{j=1}^n \cos(2\pi n_i \frac{j}{n}) = \sum_{j=1}^n \sin(2\pi n_i \frac{j}{n}) = \sum_{j=1}^n \cos(2\pi n_i \frac{j}{n}) \sin(2\pi n_{i'} \frac{j}{n}) = 0.$$

b) Montrer que les estimateurs (du maximum de vraisemblance) des paramètres sont :

$$\hat{a}_0 = \frac{1}{n} \sum_{j=1}^n Y_j$$

$$\hat{a}_i = \frac{1}{p} \sum_{j=1}^n Y_j \cos(2\pi n_i \frac{j}{n}), \quad i = 1, 2$$

$$\hat{b}_i = \frac{1}{p} \sum_{j=1}^n Y_j \sin(2\pi n_i \frac{j}{n}), \quad i = 1, 2$$

$$S^2 = \frac{1}{2p-5} \sum_{j=1}^n \left[ Y_j - \left( \hat{a}_0 + \hat{a}_1 \cos(2\pi n_1 \frac{j}{n}) + \hat{a}_2 \cos(2\pi n_2 \frac{j}{n}) + \hat{b}_1 \sin(2\pi n_1 \frac{j}{n}) + \hat{b}_2 \sin(2\pi n_2 \frac{j}{n}) \right) \right]^2$$

c) Construire le test d'hypothèse de  $a_1 = a_2 = b_1 = b_2 = 0$ , puis celui de  $a_1 = b_1 = 0$ .

d) Application numérique pour le second test de la question précédente. Cas de la note  $LA$ ,  $n_2 = 440$ ,  $n = 1000$ . Tester si la note est un  $LA$  pur (c'est-à-dire  $a_1 = b_1 = 0$ ) avec  $S_{\text{obs}}^2 = 1,5$ ,  $\hat{a}_1^{\text{obs}} = 0,52$ ,  $\hat{b}_1^{\text{obs}} = 0,98$

## 2 Régression

On considère le modèle de régression multiple :

$$Y_j = a + bx_j + cz_j + \epsilon_j, \quad j = 1 \dots 4n.$$

Où :

–  $a, b, c$  sont des paramètres inconnus.

–  $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{4n} \end{pmatrix} \sim \mathcal{N}_{4n}(0, \sigma^2 I_{4n})$ , la variance  $\sigma^2$  est inconnue.

–  $x_{4j-3} = -1, x_{4j-2} = 1, x_{4j-1} = -1, x_{4j} = 1, z_{4j-3} = 1, z_{4j-2} = 1, z_{4j-1} = -1, z_{4j} = -1,$   
( $j = 1, \dots, n$ ).

a) Estimer les paramètres  $a, b, c, \sigma^2$ . Dans la suite on note  $S^2$  l'estimateur de  $\sigma^2$ .

b) Tester l'hypothèse  $a = 0$ . Application numérique :  $n = 50, S_{\text{obs}}^2 = 2.0, \bar{Y}_{\text{obs}} = 0.17$ .

c) Tester l'hypothèse  $b = c$ .

d) Quelle est la loi de

$$\frac{1}{4n\sigma^2} \left[ \left( \sum_{j=1}^{4n} x_j \epsilon_j \right)^2 + \left( \sum_{j=1}^{4n} z_j \epsilon_j \right)^2 \right] ?$$

## 3 Régression (Bis)

On considère les modèles de régression :

$$Y_i = \alpha x_i + \varepsilon_i, \quad i = 1, \dots, n$$

$$Z_i = \beta w_i + \varepsilon'_i, \quad i = 1, \dots, n.$$

$\varepsilon_1, \dots, \varepsilon_n, \varepsilon'_1, \dots, \varepsilon'_n$  sont des suites i.i.d. indépendantes de loi  $\mathcal{N}(0, \sigma^2)$ ,  $\sigma > 0$ .

a) Sous quelle condition sur  $(x_i)_{i=1 \dots n}$  et  $(w_i)_{i=1 \dots n}$  les paramètres  $\alpha$  et  $\beta$  sont-ils estimables ?  
On supposera dans la suite que cette condition est vérifiée.

b) Estimer les paramètres  $\alpha, \beta$  et  $\sigma^2$ .

c) Tester  $\alpha = \beta$ .

d) On suppose que  $\alpha$  a pour loi a priori, la loi  $\mathcal{N}(0, \tau^2)$ ,  $\tau > 0$ . Calculer, l'estimateur bayésien de  $\alpha$ .

e) On suppose que  $\beta$  a pour loi a priori, la loi de Bernoulli de paramètre 0.5. Que signifie cette hypothèse ? Calculer, l'estimateur bayésien de  $\alpha$ .

## 4 Régression périodique

On considère, pour  $N \geq 1$ , le modèle de régression :

$$Y_j = ax_j + b \cos x_j + \epsilon_j, \quad j = -N \dots N.$$

Où :

–  $a, b$  sont des paramètres inconnus.

–  $\epsilon = \begin{pmatrix} \epsilon_{-N} \\ \vdots \\ \epsilon_N \end{pmatrix} \sim \mathcal{N}_{2N+1}(0, \sigma^2 I_{2N+1})$ , la variance  $\sigma^2$  est inconnue.

–  $x_j = \frac{j\pi}{N}$ , ( $j = -N, \dots, N$ ).

– Montrer les formules suivantes :

$$\sum_{j=1}^N j^2 = \frac{N(N+1)(2N+1)}{6}$$

$$\sum_{j=-N}^{N-1} \cos^2(x_j) = N$$

- b) Estimer les paramètres  $a, b, \sigma^2$ . Dans la suite on note  $S^2$  l'estimateur de  $\sigma^2$ .  
 c) Tester l'hypothèse  $a = 0$ .  
 d) Tester l'hypothèse  $b = 0$ .

## 5 Régression polynômiale

Soient  $p$  et  $n$  deux entiers tels que  $0 \leq p < n$ . Pour tout  $1 \leq j \leq n$  on pose  $x_j = j - \frac{n+1}{2}$ . On considère  $\psi_0, \psi_1, \dots, \psi_p$  polynômes fixés tels que :

- $\text{degré}(\psi_i) = i, \forall 0 \leq i \leq p$
- $\sum_{j=1}^n \psi_i(x_j)\psi_k(x_j) = 0, \forall 0 \leq i \neq k \leq p$

On a en vue d'étudier le modèle de régression

$$Y_j = \sum_{i=0}^p \lambda_i \psi_i(x_j) + \varepsilon_j, \quad 1 \leq j \leq n$$

où  $\varepsilon_1, \dots, \varepsilon_n$  sont indépendantes et de même loi  $\mathcal{N}(0, \sigma^2)$  et  $\lambda_0, \dots, \lambda_p, \sigma^2$  sont des paramètres inconnus.

- a) Estimer les paramètres de ce modèle.  
 b) Ecrire un test de niveau  $\alpha$  de " $\lambda_p = 0$ " contre " $\lambda_p \neq 0$ ".

Le revenu net par action de la compagnie Gillette pour les années 57 à 64 est le suivant :

Année ( $z_j$ )	:	1957	1958	1959	1960	1961	1962	1963	1964
Revenu en \$ ( $Y_j(\omega)$ )	:	0,93	0,99	1,11	1,33	1,52	1,60	1,47	1,33

On pose  $x_j = z_j - 1960,5$  puis  $\psi_0(x) = 1, \psi_1(x) = 2x$  et  $\psi_2(x) = x^2 - 21/4$ .

c) Vérifier brièvement que le choix ci-dessus de  $x_j, \psi_0, \psi_1, \psi_2$  rentre dans le cadre décrit plus haut.

d) En supposant que  $Y_j = \sum_{i=0}^2 \lambda_i \psi_i(x_j) + \varepsilon_j, \quad 1 \leq j \leq 8$  avec  $\varepsilon_1, \dots, \varepsilon_8$  indépendantes et de même loi  $\mathcal{N}(0, \sigma^2)$ , estimer  $\lambda_0, \lambda_1, \lambda_2$  et  $\sigma^2$ . Tester si  $\lambda_2 = 0$ .

e) Faire une prévision pour le revenu net par action en 1965.

N.B. Pour faciliter les calculs, on indique les valeurs numériques suivantes :

$$\sum_j \psi_1^2(x_j) = \sum_j \psi_2^2(x_j) = 168 \quad \sum_j Y_j^2(\omega) = 13,65 \quad \sum_j \psi_0(x_j)Y_j(\omega) = 10,28$$

$$\sum_j \psi_1(x_j)Y_j(\omega) = 6,86 \quad \sum_j \psi_2(x_j)Y_j(\omega) = -4,1.$$

## 6 Degré 3

Pour les données ci-dessous, ajuster le modèle :

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

Tester l'hypothèse  $\beta_2 = \beta_3$ .

X	-5	-3	-1	1	3	5
Y	13	4	3	4	10	22

## 7 Régression

Soit  $p$  un entier naturel strictement positif, on considère le modèle linéaire

$$Y_{ij} = a_i x_{ij} + b_i + \varepsilon_{ij}, \quad i = 1, 2, \quad j = 1 \dots 2p,$$

où

- $\varepsilon_{ij}$ ,  $i = 1, 2$ ,  $j = 1 \dots 2p$ , sont des variables i.i.d., gaussiennes de variance  $\sigma^2$ .
- $x_{ij} = (-1)^{i+j}$ ,  $i = 1, 2$ ,  $j = 1 \dots 2p$ .
- Les paramètres  $a_i$ ,  $b_i$   $i = 1, 2$  et  $\sigma^2$  sont inconnus.

- a) Calculer les estimateurs du maximum de vraisemblance des paramètres.
- b) Déterminer une région de confiance pour le couple de coefficients directeurs  $(a_1, a_2)$ . Pour  $p = 50$ , on suppose que l'on a observé  $S^{2\text{obs}} = 0.9$  ( $S^2$  est l'estimateur du maximum de vraisemblance de  $\sigma^2$ ), et  $(\sum_{j=1}^{2p} x_{1j} Y_{1j})^{\text{obs}} = 2.5$   $(\sum_{j=1}^{2p} x_{2j} Y_{2j})^{\text{obs}} = 1.4$ . Tracer, pour ces observations, la région de confiance obtenue dans le plan  $(a_1, a_2)$ .
- c) Tester l'égalité des coefficients directeurs  $a_1 = a_2$ .

Application numérique : reprendre les valeurs numériques données à la question précédente.

- d) Tester l'égalité des ordonnées à l'origine  $b_1 = b_2$ .

Application numérique : on reprend les valeurs de  $p$  et de  $S^2$  de la question b) et l'on suppose que  $Y_1^{\text{obs}} = (1/(2p) \sum_{j=1}^{2p} Y_{1j})^{\text{obs}} = 5.3$   $Y_2^{\text{obs}} = (1/(2p) \sum_{j=1}^{2p} Y_{2j})^{\text{obs}} = 4.95$ .

- e) On suppose que  $a_1 = a_2 = a$  et  $b_1 = b_2 = 0$ . Calculer l'estimateur de Bayes de  $a$ , si la loi a priori est la loi  $\mathcal{N}(m, \xi^2)$  ( $m \in \mathbb{R}, \xi > 0$ ).

## 8 Modèle linéaire et test d'hypothèses non linéaires

Soit  $\varepsilon_{ij}$ ,  $i = 1, \dots, k$   $j = 1 \dots n$ , des variables i.i.d., de loi normale centrée réduite. Soit  $m_1, \dots, m_k \in \mathbb{R}$ , on pose  $Y_{ij} = m_i + \varepsilon_{ij}$   $i = 1, \dots, k$   $j = 1 \dots n$ .

- a) Quels sont les estimateurs du maximum de vraisemblance  $\hat{m}_i$  de  $m_i$  ( $i = 1, \dots, k$ ) ?
- b) On se propose maintenant de construire une statistique de test pour les hypothèses non linéaires :

$$H_0 \quad \sum_{i=1}^k m_i^2 = 1 \quad H_1 \quad \sum_{i=1}^k m_i^2 \neq 1.$$

Pour cela, on se propose d'utiliser la statistique de test

$$T_n = \sqrt{n} \left( \sum_{i=1}^k \hat{m}_i^2 - 1 \right).$$

Montrer que  $T_n$  peut s'écrire sous la forme suivante

$$T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^k (\xi_i^n)^2 + 2 \sum_{i=1}^k (\xi_i^n) m_i + \sqrt{n} \left( \sum_{i=1}^k m_i^2 - 1 \right),$$

où  $(\xi_i^n)_{i=1 \dots k}$  sont i.i.d. de loi normale standard.

- c) On suppose que  $H_0$  est vérifiée. Montrer que  $\frac{1}{\sqrt{n}} \sum_{i=1}^k (\xi_i^n)^2$  converge en probabilité vers 0. En déduire que  $T_n$  converge en loi. Préciser la loi limite.
- d) On suppose que  $H_1$  est vérifiée. En utilisant la question précédente, montrer que presque sûrement  $\lim_{n \rightarrow \infty} |T_n| = +\infty$ .
- e) Déduire des questions précédentes une procédure de test pour décider entre  $H_0$  et  $H_1$ .